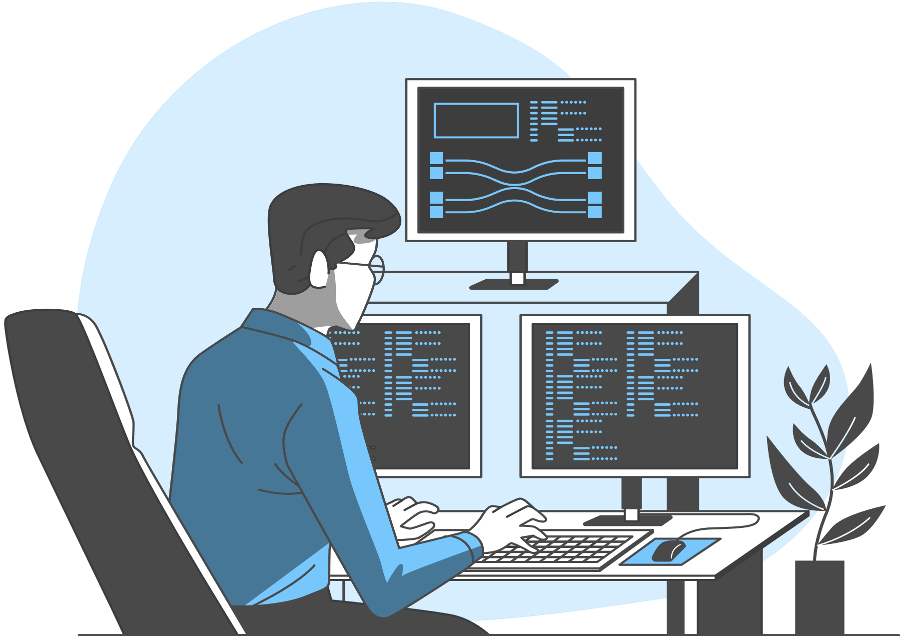
Mining Airbnb Rental Data
Group 13



Group Members:

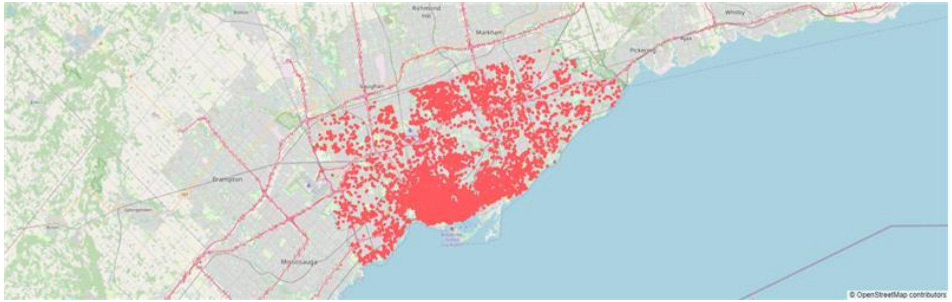
Ahmed AbdElaziz (21amga)
Karim Mohamed (21kgmm)
Sara Elfetiany (21sama)

01

Datastes Description

Datastes Description

- We found many datasets for Airbnb, but we chose a Toronto dataset to deal with
- We have 4 Datasets for Toronto, but we found that we can provide our questions with only two files(Listings, Calendar)



Datastes Description

Airbnb provides open data that quantifies the impact of short-term rentals on housing and residential communities; and also provides a platform to support advocacy for policies to protect our cities from the impacts of short-term rentals. By analyzing publicly available information about a city's Airbnb's listings, Inside Airbnb provides filters and key metrics so we can see how Airbnb is being used to compete with the residential housing market.

The dataset comprises of four main tables:

1. Listings - Detailed listings data showing 74 attributes for each of the listings. Some of the attributes used in the analysis are price, longitude, latitude, listing type, neighborhood, ratings among others, and other attributes related to place. – dimensions: (15261, 74)

2. Reviews - Detailed reviews given by the guests with 6 attributes. Key attributes include date, listing id, reviewer id and comment. – dimensions: (400423, 6)

3. Calendar - Provides details about booking for the next year by listing. 7 attributes in total including listing id, date, available and price. – dimensions: (5569545, 7)

4. Neighborhoods – It doesn't have important information to our analysis so we wouldn't focus on it

02

Data Preprocessing

Data Preprocessing

01

Removing Duplicates

Check if there is any duplicate rows and remove them

02

Handling Missing Values

- Remove columns that has nulls more than 30%
- Remove some records with null values
- Using imputation (IterativeImputer, SimpleImputer)

03

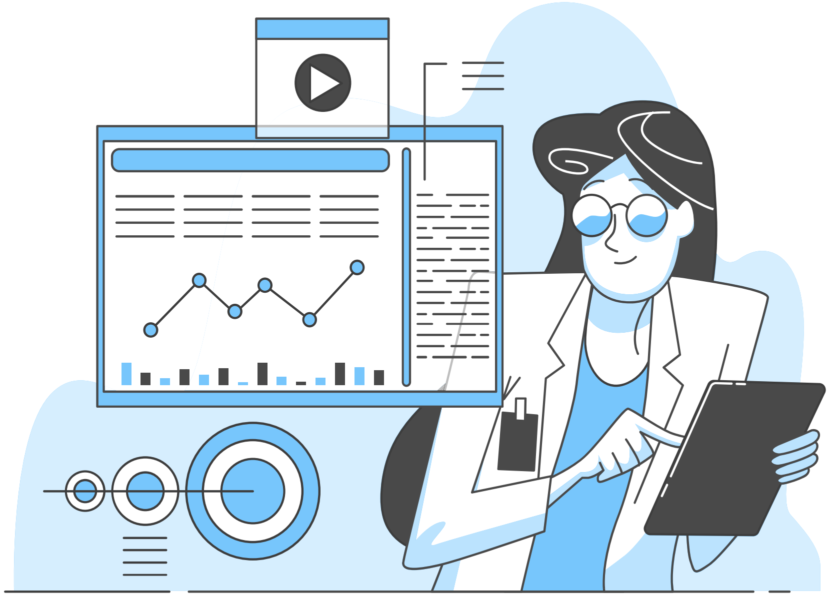
Handling Outliers

- Replacing by the mean or median
- Removing outliers

04

Handling Datatyps

Convert categorical features to numerical features



03

Analysis Questions

Hypothes Test

The average of listings price with a rating of 4.5 or higher are more expensive than those with a lower rating!

Null Hypothesis & Alternative Hypothesis

- The null hypothesis (H0) is: The rating of listings that higher than or equal 4.5 has no effect on increasing the price

- The alternative hypothesis (H1) is: The average price of listings with a rating of 4.5 or higher are more expensive than those with a lower rating

Motivation

This will be useful for people who are looking for cheap prices for real estate with a good rating.

...

Approach

- The statistical test we used is T-test (using stats.ttest_ind)
- We can use this test, if we observe two independent samples from the same or different population. The test measures whether the average (expected) value differs significantly across samples. If we observe a large p-value, for example larger than 0.05, then we cannot reject the null hypothesis of identical average scores. If the p-value is smaller than the threshold, then we reject the null hypothesis of equal averages.

Findings

- The result of hypothesis test

```
0.0011696354362489883
Reject null hypothesis
```

- The hypothesis test failed which mean that rating of listing affect its price and the average of listings price with a rating of 4.5 or higher are more expensive than those with a lower rating!

Regression Analysis

Can we predict the price of the listings based on available features?

Motivation

It will benefit for customers to find the best listings price based on their features

...

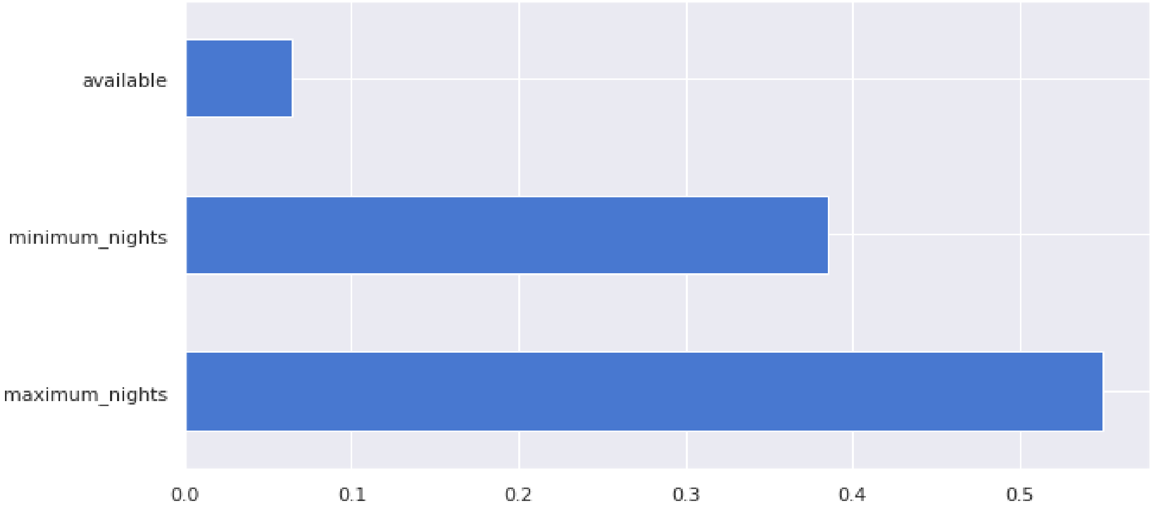
Approach

Step 1
Making feature selection based on the price

Step 2
Selecting the top 20 features correlated with price

Step 3
Train the model & Make prediction based on these selected features

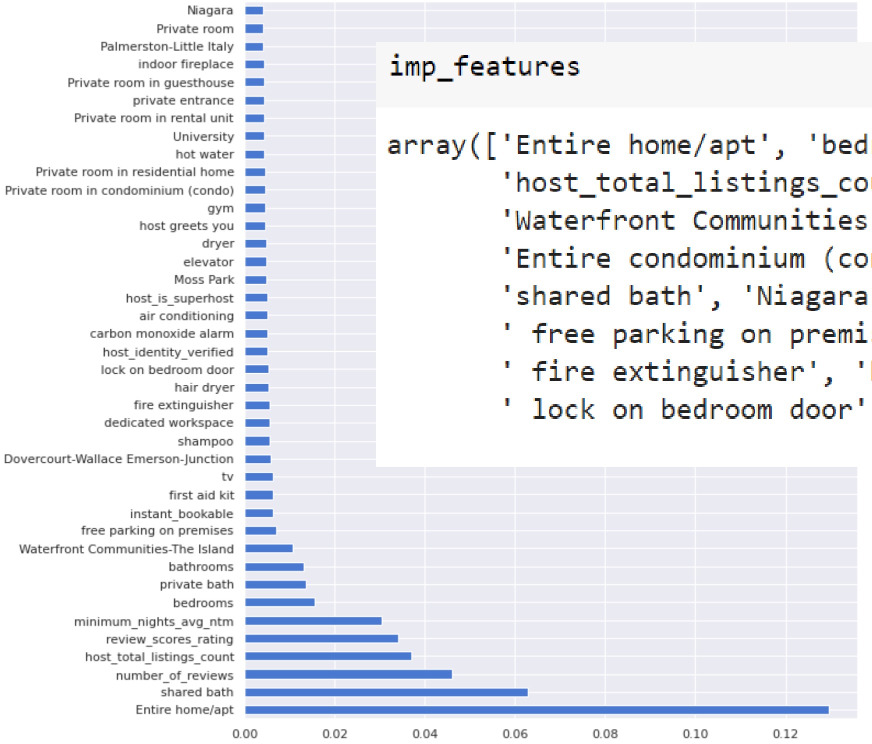
Results – Feature Selection (Calendar)




```
imp_features
```

```
array(['maximum_nights', 'minimum_nights'], dtype=object)
```

Results – Feature Selection (Listings)




```
imp_features
```

```
array(['Entire home/apt', 'bedrooms', 'minimum_nights_avg_ntm',
       'host_total_listings_count', 'bathrooms',
       'Waterfront Communities-The Island', 'number_of_reviews',
       'Entire condominium (condo)', 'review_scores_rating',
       'shared bath', 'Niagara', 'Entire loft', 'instant_bookable',
       ' free parking on premises', ' tv', ' dedicated workspace',
       ' fire extinguisher', 'host_identity_verified', ' first aid kit',
       ' lock on bedroom door'], dtype=object)
```

Findings

- Model result (Calendar)

  Results of sklearn.metrics:
  MAE: 44.97802304637844
  MSE: 3140.5343929953237
  RMSE: 56.04047102760043
  R-Squared: 0.07147009133589355

- Model result (Listings)

  Results of sklearn.metrics:
  MAE: 30.317636740958033
  MSE: 1654.457445784906
  RMSE: 40.67502238210701
  R-Squared: 0.47759214502274794

Predictive Analysis

Can we predict the price level of the listings after converting the prices to three levels(categories): low, medium, high. according to: ‘prices that are less or equal to 25% as (low), 25% to 75% as (medium), and 75% or higher as (high)’ ?

Motivation

This method will make it easier for user to find the suitable listings based on the category of the listings price (the price range)

...

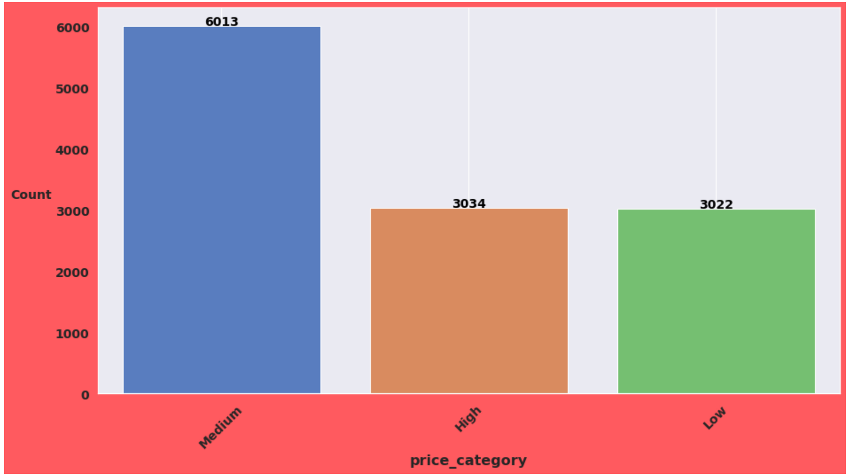
Approach

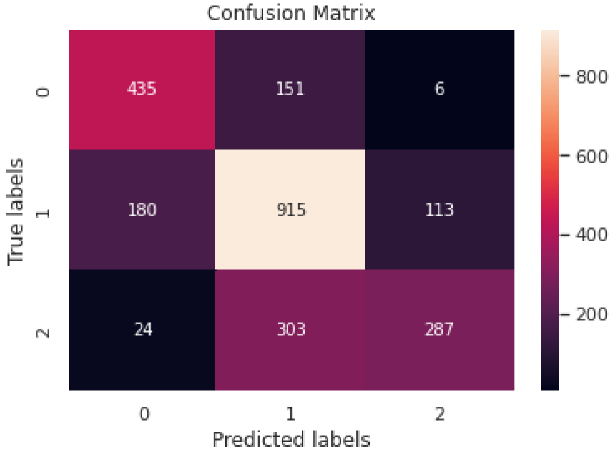
Step 1
Create price category column that contain only three values: Low, Medium, High

Step 2
Convert categorical column to numerical (Low = 0, Medium = 1, High = 2)

Step 3
Make prediction based on the price categories after making feature selection

Results (Listings)



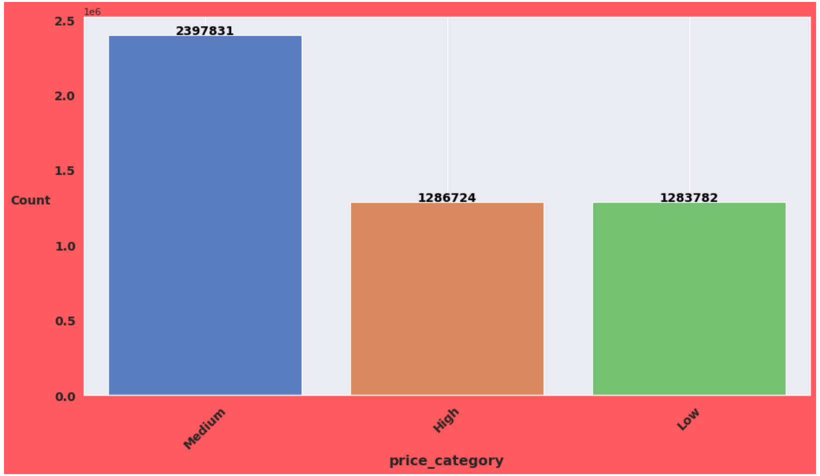


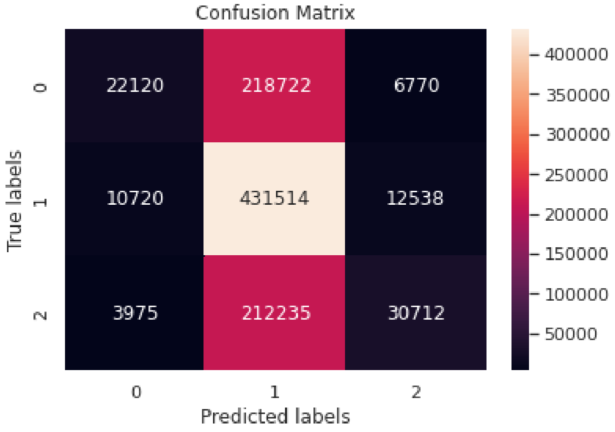
Results of sklearn.metrics:
MAE: 0.33429991714995855
MSE: 0.3591549295774648
RMSE: 0.5992953608843179
R-Squared: 0.2809749896935517

Classification Report is :

	precision	recall	f1-score	support
0	0.68	0.73	0.71	592
1	0.67	0.76	0.71	1208
2	0.71	0.47	0.56	614
accuracy			0.68	2414
macro avg	0.69	0.65	0.66	2414
weighted avg	0.68	0.68	0.67	2414

Results (Calendar)





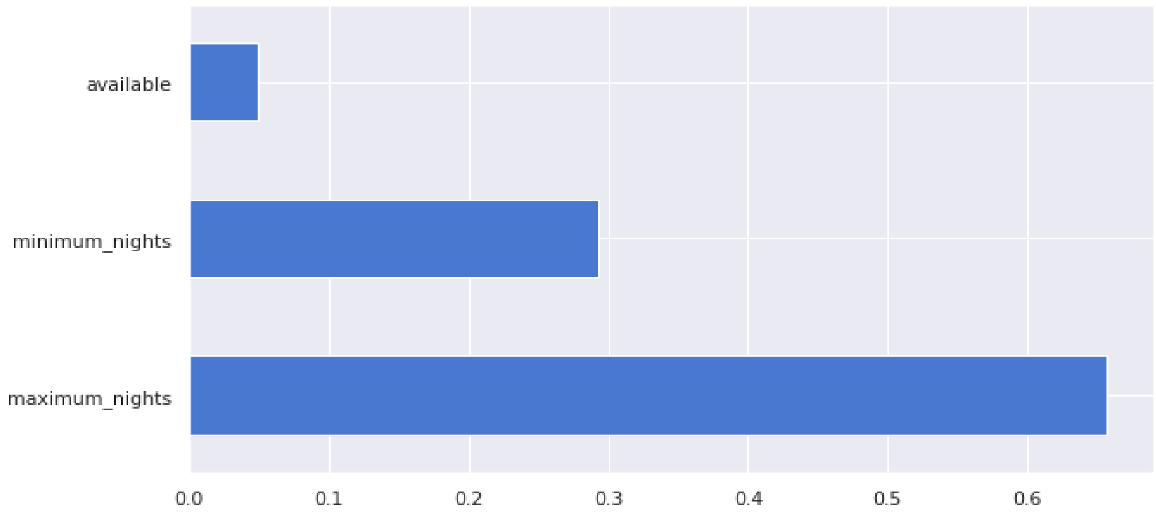

```
Results of sklearn.metrics:
MAE: 0.5011081779742254
MSE: 0.523745767961O157
RMSE: 0.7237028174333825
R-Squared: -0.005381842751755128
```

```
Classification Report is :
              precision    recall  f1-score   support

           0       0.60      0.09      0.16    247612
           1       0.50      0.95      0.66    454772
           2       0.61      0.12      0.21    246922

    accuracy                           0.51    949306
   macro avg       0.57      0.39      0.34    949306
weighted avg       0.56      0.51      0.41    949306
```

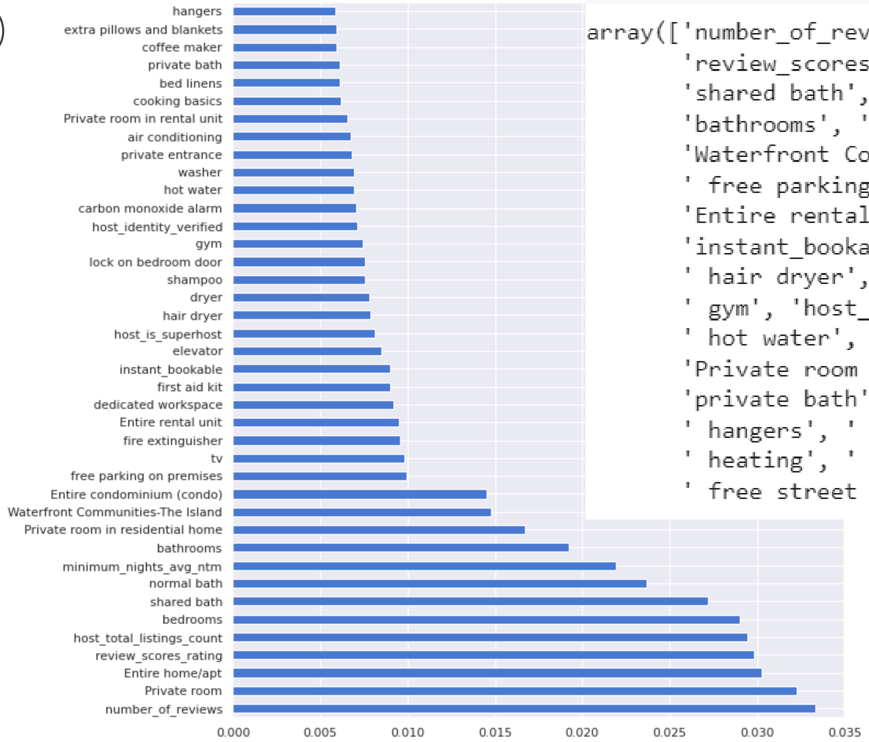
Results – Feature Selection (Calendar)




```
imp_features
```

```
array(['maximum_nights', 'minimum_nights'], dtype=object)
```

Results – Feature Selection (Listings)




```
imp_features
```

```
array(['number_of_reviews', 'Private room', 'Entire home/apt',
       'review_scores_rating', 'host_total_listings_count', 'bedrooms',
       'shared bath', 'normal bath', 'minimum_nights_avg_ntm',
       'bathrooms', 'Private room in residential home',
       'Waterfront Communities-The Island', 'Entire condominium (condo)',
       ' free parking on premises', ' tv', ' fire extinguisher',
       'Entire rental unit', ' dedicated workspace', ' first aid kit',
       'instant_bookable', ' elevator', 'host_is_superhost',
       ' hair dryer', ' dryer', ' shampoo', ' lock on bedroom door',
       ' gym', 'host_identity_verified', ' carbon monoxide alarm',
       ' hot water', ' washer', ' private entrance', ' air conditioning',
       'Private room in rental unit', ' cooking basics', ' bed linens',
       'private bath', ' coffee maker', ' extra pillows and blankets',
       ' hangers', ' iron', '[shampoo', ' host greets you', ' dishwasher',
       ' heating', ' smoke alarm', ' cable tv', 'Entire residential home',
       ' free street parking', ' luggage dropoff allowed'], dtype=object)
```

Findings

- We noticed that the number of hosts whose prices are medium is twice the number of hosts whose prices are high and cheap.
- We conclude from this that most of the prices in Toronto are medium prices

04

Limitation

Projrct Limitations

01

Huge Data

We have encountered limitations in the large size of the calendar data, although the size of the data was smaller after preprocessing, but we had almost 5 million records, which make us to train the Random Forest Classifier on 50% of the data not 85% or 90% as usual.
We also faced a problem with the large number of features for Listing dataset, which were 74 features, and there was not enough information to facilitate the preprocessing of them even on the Airbnb official website .

02

Generalization

This analysis wouldn’t be generalized to locations other than Toronto. Also, as the time going by, this analysis may not be very effective

05

Conclusion

Conclusion

- What we discovered is that the price prediction will be more accurate using 20 of the listings features, and it will not be as accurate as the features of the calendar file.

- We noticed that the number of hosts whose prices are medium is twice the number of hosts whose prices are high and cheap.

- rating of listing affect its price and the average of listings price with a rating of 4.5 or higher are more expensive than those with a lower rating!

Thank you for listening!